IV107 Bioinformatika I

Přednáška 5

Katedra informačních technologií
Masarykova Univerzita Brno

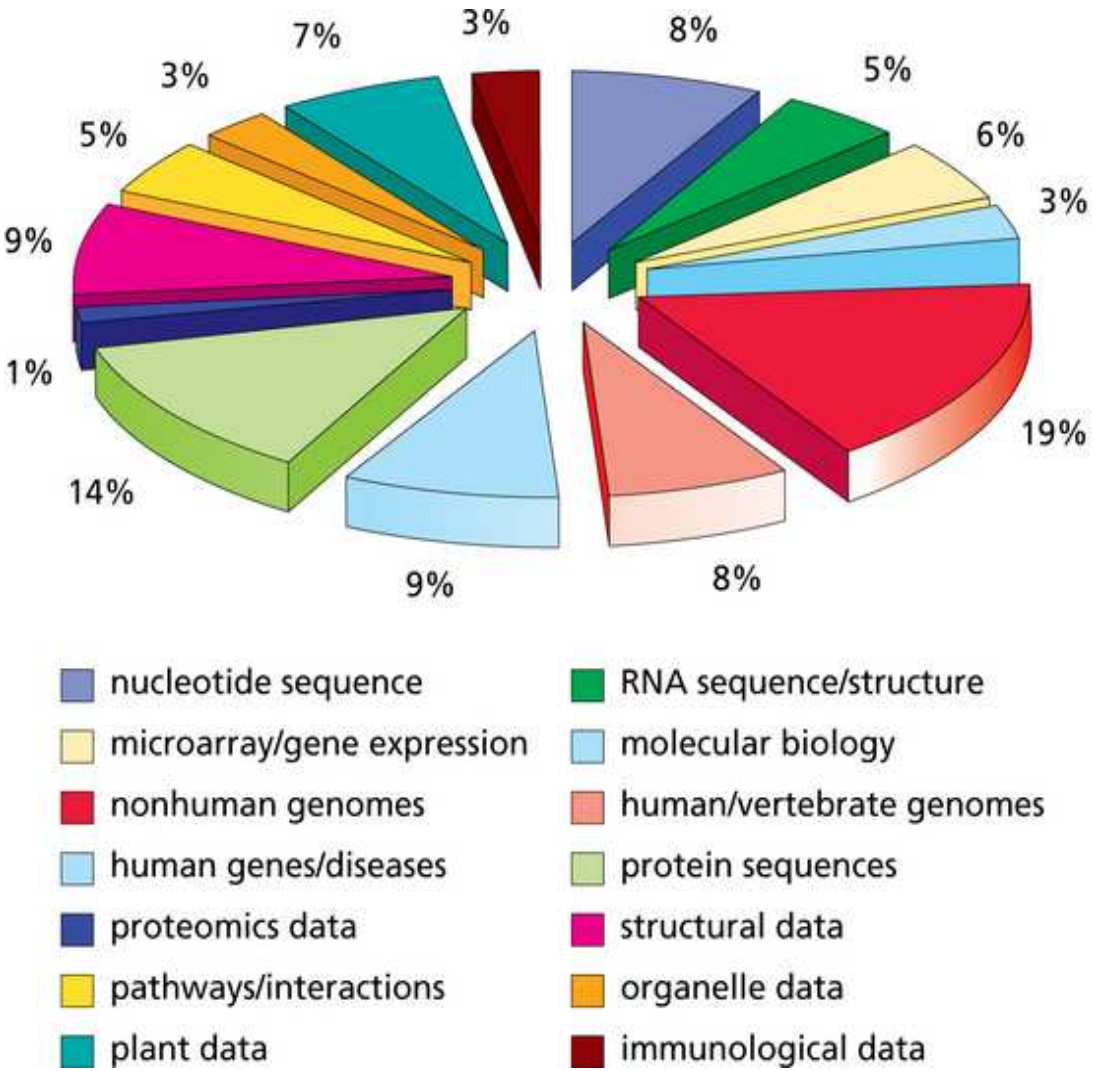
Jaro 2019

Předchozí týden

- Struktura genu
  - prokaryotického
  - eukaryotického
- Porovnání sekvencí
  - globální (Needleman–Wunsch)
  - semi-globální
  - lokální (Smith–Waterman)

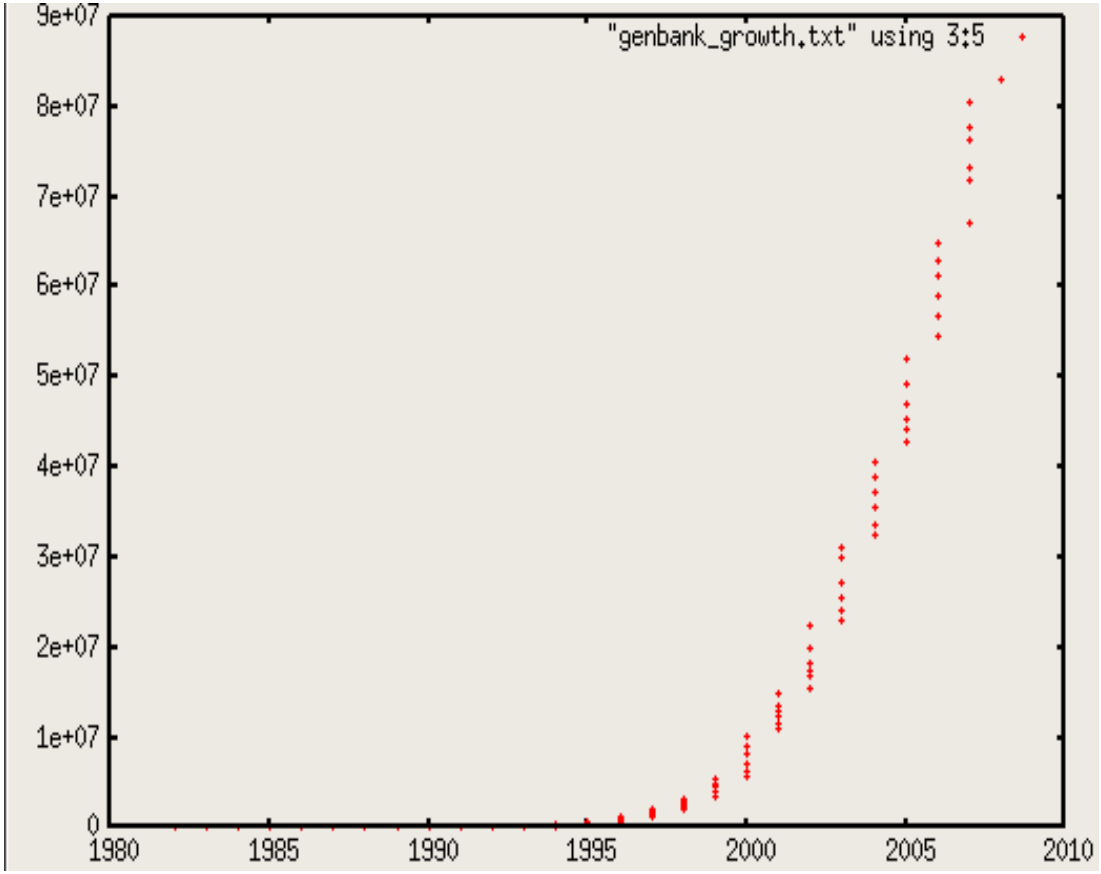
Outline

Typy dat v databázích



http://www.agr.kuleuven.ac.be/vak

Nárůst databáze GenBank



GenBank

Genetic Sequence Data Bank
August 2009
NCBI-GenBank Flat File Release 164.0
National Center for Biotechnology Information

- 106533156756 bp
- 108431692 sekv.

ftp://http://www.ncbi.nlm.nih.gov/genbank/

GenBank

NCBI-GenBank Flat File Release 232.0
June 15 2019
Distribution Release Notes

- 329 835 282 370 bp
- 213 383 758 sekv.

ftp://ftp.ncbi.nlm.nih.gov/genbank/

Součásti databáze GenBank

- INV, VRT, MAM, PLN, PRI, ROD, BCT, VRL
- PAT (Patents)
- HTGS (High Throughput Genomic Sequences)
- GSS (Genome Survey Sequences)
- ETS (Expressed Sequence Tags)
- STS (Sequence Tagged Sites)
- WGS (Whole Genome Shotgun)

Příklad záznamu v databázi GenBank


```
LOCUS       SCU49845     5028 bp    DNA
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene,
            Axl2p
            (AXL2) and Rev7p (REV7) genes, complete
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomy
           Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; S
```

Vyhledávání v sekvenčních databázích

- textové (klíčová slova)
- sekvenční (BLAST)

GenBank

Uniprot
September, 2019
UniprotKB release 2019_08
The UniProt consortium: European Bioinformatics Institute (EBI), Swiss Institute of Bioinformatics (SIB) and Protein Information Resource (PIR)

- 560,823 (SwissProt)
- 171,501,488 (TrEMBL)
- 37,597,356 (UniRef50)

Release 2019_08 of 18-Sep-19 of UniProtKB/Swiss-Prot contains 560823 sequence entries, comprising 201585439 amino acids abstracted from 268349 references.
http://expasy.org/sprot/

Příklad záznamu v databázi UniProt

Entry information

Entry name	**LMO7_HUMAN**
Primary accession number	**Q8WWI1**
Secondary accession numbers	O15462 O95346 Q9UKC1 Q9UQM5 Q9Y6A7
Integrated into Swiss-Prot on	March 15, 2004
Sequence was last modified on	March 15, 2004 (Sequence version 2)
Annotations were last modified on	July 25, 2006 (Entry version 39)

Name and origin of the protein

Protein name	**LIM domain only protein 7**
Synonyms	**LOMP** **F-box only protein 20**
Gene name	**Name:** **LMO7** Synonyms: FBX20, FBXO20, KIAA0858
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

References

[1] NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 3), AND TISSUE SPECIFICITY.
TISSUE=Brain, and Peripheral blood leukocyte;
DOI=10.1007/s00439-001-0646-6; PubMed=11935316 [NCBI, ExPASy, EBI, Israel, Japan]
Rozenblum E., Vahteristo P., Sandberg T., Bergthorsson J.T., Syrjakoski K., Weaver D., Haraldsson K., Johannsdottir H.K., Vehmanen P., Nigam S., Golberger N., Robbins C., Pak E., Dutra A., Gillander E., Stephan D.A., Bailey-Wilson J., Juo S.-H.H., Kainu T.,  , Kallioniemi O.-P.;
"A genomic map of a 6-Mb region at 13q21-q22 implicated in cancer development: identification and characterization of candidate genes.";
Hum. Genet. 110:111-121(2002).

http://www.uniprot.org/

Příklad záznamu v databázi UniProt

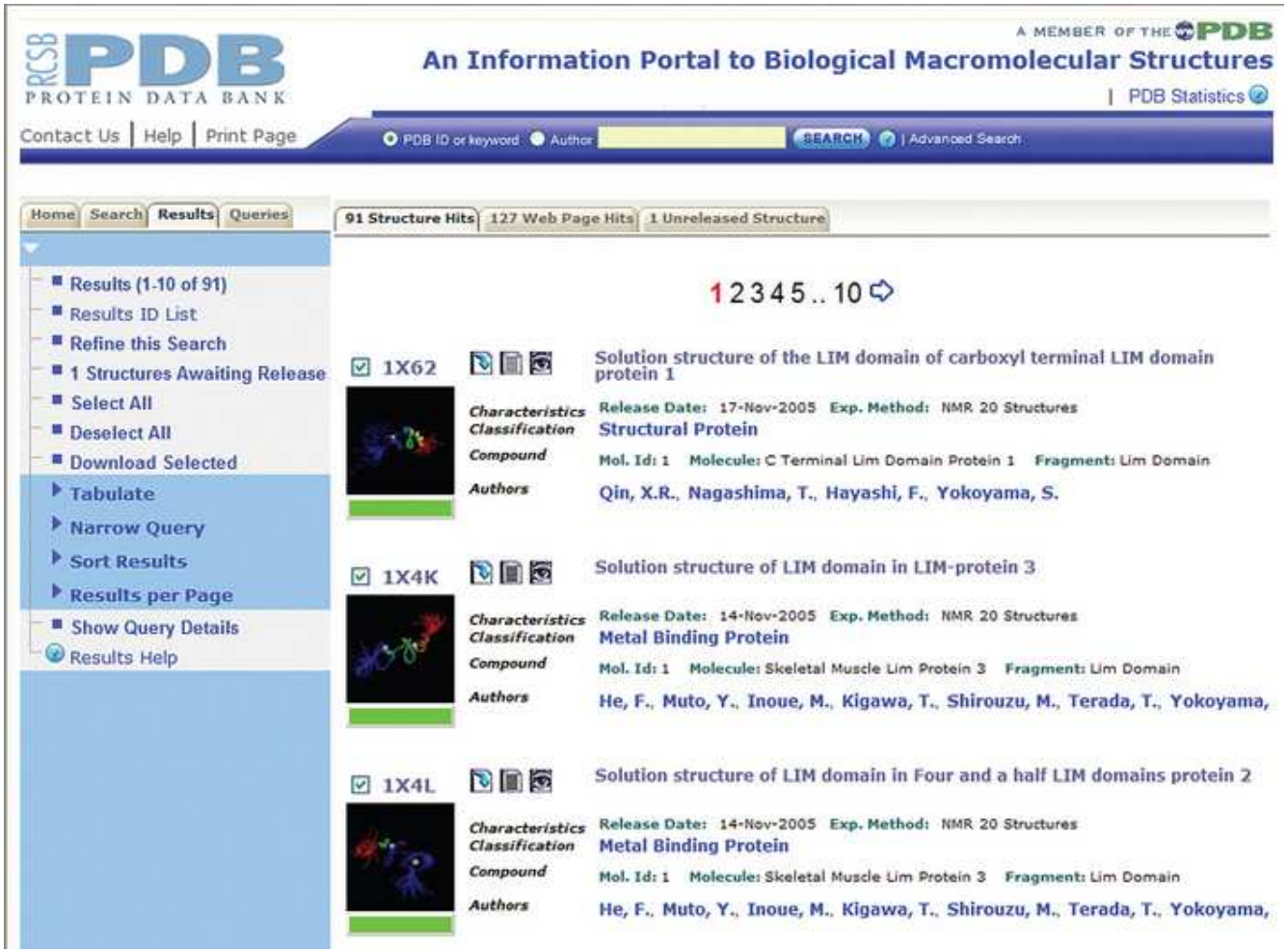
Key	From	To	Length	Description	FTId
CHAIN	1	1683	1683	LIM domain only protein 7.	PRO_0000075824
DOMAIN	54	168	115	CH.	
DOMAIN	1042	1128	87	PDZ.	
DOMAIN	1612	1678	67	LIM zinc-binding.	

```
        10         20         30         40         50         60
MKKIRICHIF TFYSWMSYDV LFQRTELGAL EIWRQLICAH VCICVGWLYL RDRVCSKKDI

        70         80         90        100        110        120
ILRTEQNSGR TILIKAVTEK NFETKDFRAS LENGVLLCDL INKLKPGVIK KINRLSTPIA

       130        140        150        160        170        180
GLDNINVFLK ACEQIGLKEA QLFHPGDLQD LSNRVTVKQE ETDRRVKNVL ITLYWLGRKA
```

PDB


RCSB PDB
PROTEIN DATA BANK
A MEMBER OF THE PDB
An Information Portal to Biological Macromolecular Structures
PDB Statistics
Contact Us | Help | Print Page
PDB ID or keyword Author SEARCH Advanced Search
Home Search Results Queries
91 Structure Hits 127 Web Page Hits 1 Unreleased Structure
Results (1-10 of 91)
Results ID List
Refine this Search
1 Structures Awaiting Release
Select All
Deselect All
Download Selected
Tabulate
Narrow Query
Sort Results
Results per Page
Show Query Details
Results Help
1 2 3 4 5 .. 10
1X62
Solution structure of the LIM domain of carboxyl terminal LIM domain protein 1
Characteristics Release Date: 17-Nov-2005 Exp. Method: NMR 20 Structures
Classification Structural Protein
Compound Mol. Id: 1 Molecule: C Terminal Lim Domain Protein 1 Fragment: Lim Domain
Authors Qin, X.R., Nagashima, T., Hayashi, F., Yokoyama, S.
1X4K
Solution structure of LIM domain in LIM-protein 3
Characteristics Release Date: 14-Nov-2005 Exp. Method: NMR 20 Structures
Classification Metal Binding Protein
Compound Mol. Id: 1 Molecule: Skeletal Muscle Lim Protein 3 Fragment: Lim Domain
Authors He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,
1X4L
Solution structure of LIM domain in Four and a half LIM domains protein 2
Characteristics Release Date: 14-Nov-2005 Exp. Method: NMR 20 Structures
Classification Metal Binding Protein
Compound Mol. Id: 1 Molecule: Skeletal Muscle Lim Protein 3 Fragment: Lim Domain
Authors He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,

Záznam v PDB


```
HEADER    HYDROLASE(O-GLYCOSYL)                   20-JAN-92   1HEW      1HEW   2
COMPND    LYSOZYME (E.C.3.2.1.17) COMPLEXED WITH THE INHIBITOR          1HEW   3
COMPND   2 TRI-N-ACETYLCHITOTRIOSE                                      1HEW   4
SOURCE    HEN (GALLUS GALLUS) EGG WHITE                                 1HEW   5
AUTHOR    J.C.CHEETHAM,P.J.ARTYMIUK,D.C.PHILLIPS                        1HEW   6
REVDAT   1   31-JAN-94 1HEW    0                                        1HEW   7
JRNL        AUTH   J.C.CHEETHAM,P.J.ARTYMIUK,D.C.PHILLIPS              1HEW   8
JRNL        TITL   REFINEMENT OF AN ENZYME COMPLEX WITH INHIBITOR       1HEW   9
JRNL        TITL 2 BOUND AT PARTIAL OCCUPANCY. HEN EGG-WHITE            1HEW  10
JRNL        TITL 3 LYSOZYME AND TRI-N-ACETYLCHITOTRIOSE AT 1.75         1HEW  11
JRNL        TITL 4 ANGSTROMS RESOLUTION                                 1HEW  12
JRNL        REF    J.MOL.BIOL.                  V. 224   613 1992      1HEW  13
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836                  070  1HEW  14
REMARK   1                                                            1HEW  15
REMARK   1 REFERENCE 1                                                1HEW  16
REMARK   1  AUTH   L.N.JOHNSON,J.C.CHEETHAM,P.J.MC*LAUGHLIN,         1HEW  17
REMARK   1  AUTH 2 K.R.ACHARYA,D.BARFORD,D.C.PHILLIPS                 1HEW  18
REMARK   1  TITL   PROTEIN-OLIGOSACCHARIDE INTERACTIONS: LYSOZYME,    1HEW  19
REMARK   1  TITL 2 PHOSPHORYLASE, AMYLASES                            1HEW  20
REMARK   1  REF    CURR.TOP.MICROBIOL.IMMUNOL.   V. 139    81 1988    1HEW  21
REMARK   1  REFN   ASTM CTMIA3  GW ISSN 0070-217X                761  1HEW  22
```

Záznam v PDB


```
REMARK   5 THE THREE SUGAR UNITS OF THE INHIBITOR MOLECULE ARE BOUND    1HEW  56
REMARK   5 IN THE UPPER THREE SITES (A TO C) OF THE LYSOZYME ACTIVE     1HEW  57
REMARK   5 SITE CLEFT.  NAG MOLECULES, NUMBERED 203, 202, AND 201, ARE  1HEW  58
REMARK   5 BOUND IN SITES A, B, AND C, RESPECTIVELY.                    1HEW  59
SEQRES   1    129  LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS  1HEW  60
SEQRES   2    129  ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY  1HEW  61
SEQRES   3    129  ASN TRP VAL CYS ALA ALA LYS PHE GLU SER ASN PHE ASN  1HEW  62
SEQRES   4    129  THR GLN ALA THR ASN ARG ASN THR ASP GLY SER THR ASP  1HEW  63
SEQRES   5    129  TYR GLY ILE LEU GLN ILE ASN SER ARG TRP TRP CYS ASN  1HEW  64
SEQRES   6    129  ASP GLY ARG THR PRO GLY SER ARG ASN LEU CYS ASN ILE  1HEW  65
SEQRES   7    129  PRO CYS SER ALA LEU LEU SER SER ASP ILE THR ALA SER  1HEW  66
SEQRES   8    129  VAL ASN CYS ALA LYS LYS ILE VAL SER ASP GLY ASN GLY  1HEW  67
SEQRES   9    129  MET ASN ALA TRP VAL ALA TRP ARG ASN ARG CYS LYS GLY  1HEW  68
SEQRES  10    129  THR ASP VAL GLN ALA TRP ILE ARG GLY CYS ARG LEU      1HEW  69
HET    NAG    201      15     N-ACETYL-D-GLUCOSAMINE                    1HEW  70
HET    NAG    202      14     N-ACETYL-D-GLUCOSAMINE                    1HEW  71
HET    NAG    203      14     N-ACETYL-D-GLUCOSAMINE                    1HEW  72
FORMUL   2  NAG    3(C8 H15 N1 O6)                                       1HEW  73
```

Záznam v PDB


```
HELIX    1   A ARG      5  HIS     15  1                                1HEW  75
HELIX    2   B LEU     25  GLU     35  1                                1HEW  76
HELIX    3   C CYS     80  LEU     84  5                                1HEW  77
HELIX    4   D THR     89  ILE     98  1                                1HEW  78
HELIX    5   E VAL    109  ASN    113  1                                1HEW  79
SHEET    1  S1 2 LYS     1  PHE     3  0                                1HEW  80
SHEET    2  S1 2 PHE    38  THR    40 -1  N  THR    40   O  LYS     1   1HEW  81
SHEET    1  S2 3 ALA    42  ASN    46  0                                1HEW  82
SHEET    2  S2 3 SER    50  GLY    54 -1  O  SER    50   N  ASN    46   1HEW  83
SHEET    3  S2 3 GLN    57  SER    60 -1  O  ILE    58   N  TYR    53   1HEW  84
TURN     1  T1 MET    12  HIS    15     TYPE III                        1HEW  85
TURN     2  T2 LYS    13  GLY    16     TYPE I                          1HEW  86
TURN     3  T3 LEU    17  TYR    20     TYPE II                         1HEW  87
TURN     4  T4 ASN    19  GLY    22     DISTORTED TYPE II               1HEW  88
TURN     5  T5 TYR    20  TYR    23     TYPE I'                         1HEW  89
TURN     6  T6 SER    24  ASN    27     TYPE III                        1HEW  90
TURN     7  T7 LEU    25  TRP    28     TYPE III                        1HEW  91
TURN     8  T8 SER    36  ASN    39     TYPE III'                       1HEW  92
```

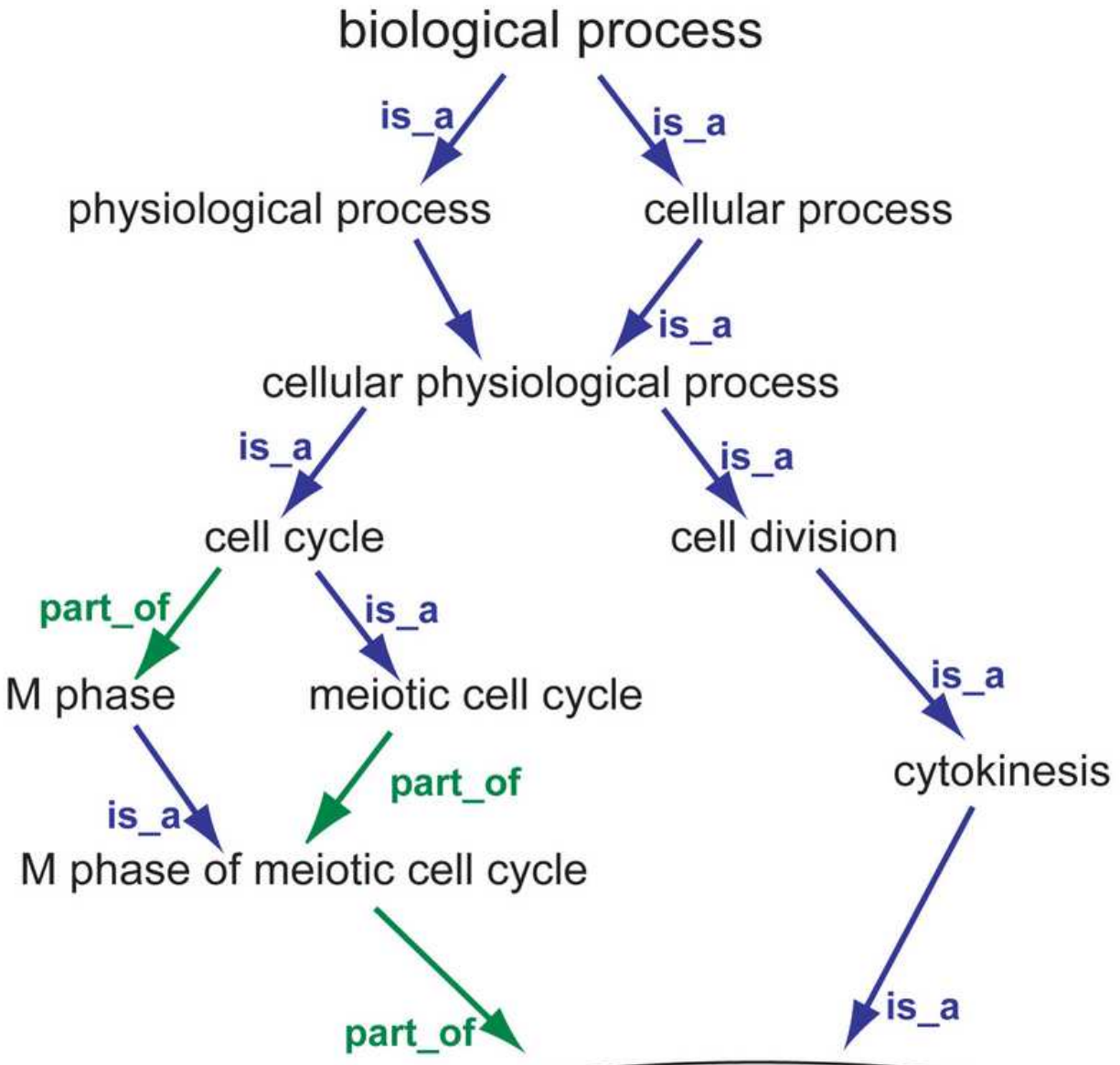
Záznam v PDB

```
CRYST1   78.860   78.860   38.250  90.00  90.00  90.00 P 43 21 2     8  1HEW 113
ORIGX1      1.000000  0.000000  0.000000        0.00000                      1HEW 114
ORIGX2      0.000000  1.000000  0.000000        0.00000                      1HEW 115
ORIGX3      0.000000  0.000000  1.000000        0.00000                      1HEW 116
SCALE1      0.012681  0.000000  0.000000        0.00000                      1HEW 117
SCALE2      0.000000  0.012681  0.000000        0.00000                      1HEW 118
SCALE3      0.000000  0.000000  0.026144        0.00000                      1HEW 119
ATOM      1  N   LYS     1       3.398   9.981  10.408  1.00 30.48           1HEW 120
ATOM      2  CA  LYS     1       2.459  10.365   9.364  1.00 28.03           1HEW 121
ATOM      3  C   LYS     1       2.458  11.880   9.149  1.00 21.93           1HEW 122
ATOM      4  O   LYS     1       2.481  12.672  10.100  1.00 14.10           1HEW 123
ATOM      5  CB  LYS     1       1.026   9.935   9.695  1.00 30.54           1HEW 124
ATOM      6  CG  LYS     1       0.028  10.169   8.558  1.00 37.93           1HEW 125
ATOM      7  CD  LYS     1      -1.415  10.089   9.048  1.00 33.23           1HEW 126
ATOM      8  CE  LYS     1      -2.357  10.822   8.082  1.00 32.17           1HEW 127
ATOM      9  NZ  LYS     1      -3.661  10.090   8.025  1.00 31.92           1HEW 128
ATOM     10  N   VAL     2       2.429  12.232   7.880  1.00 17.30           1HEW 129
ATOM     11  CA  VAL     2       2.395  13.653   7.465  1.00 14.47           1HEW 130
ATOM     12  C   VAL     2       0.977  13.868   6.903  1.00 17.58           1HEW 131
ATOM     13  O   VAL     2       0.642  13.368   5.826  1.00 32.65           1HEW 132
ATOM     14  CB  VAL     2       3.533  14.012   6.536  1.00 22.88           1HEW 133
```

Gene Ontology

- Funkce genů a proteinů zjišťujeme experimentálně
- Slovní popis není jednoznačný
  - syntéza proteinů
  - syntéza polypeptidů
  - translace
  - aktivita ribozomů
- Ontologie je způsob jak do používaných termínů vnést systém

Gene Ontology



Gene Ontology

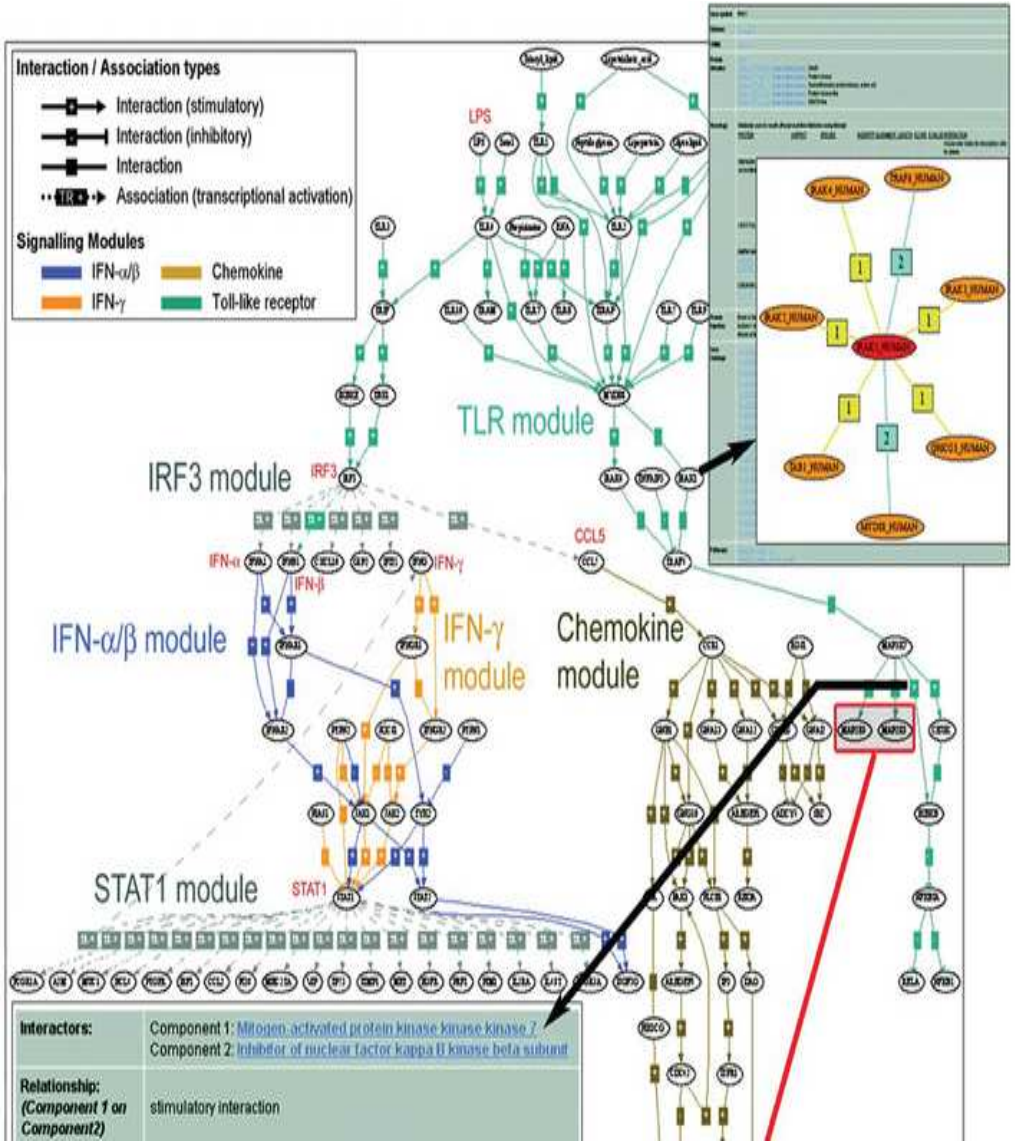
- Molekulární proces
  - katalytická aktivita
  - transport
  - intermolekulární vazba
- Biologický proces
  - přenos signálu
  - aktivace imunitního sytému
  - regulace genů
- Buněčná složka
  - buněčné jádro
  - plazmatická membrána

Gene Ontology - kódy zdroje dat

Curator-assigned Evidence Codes

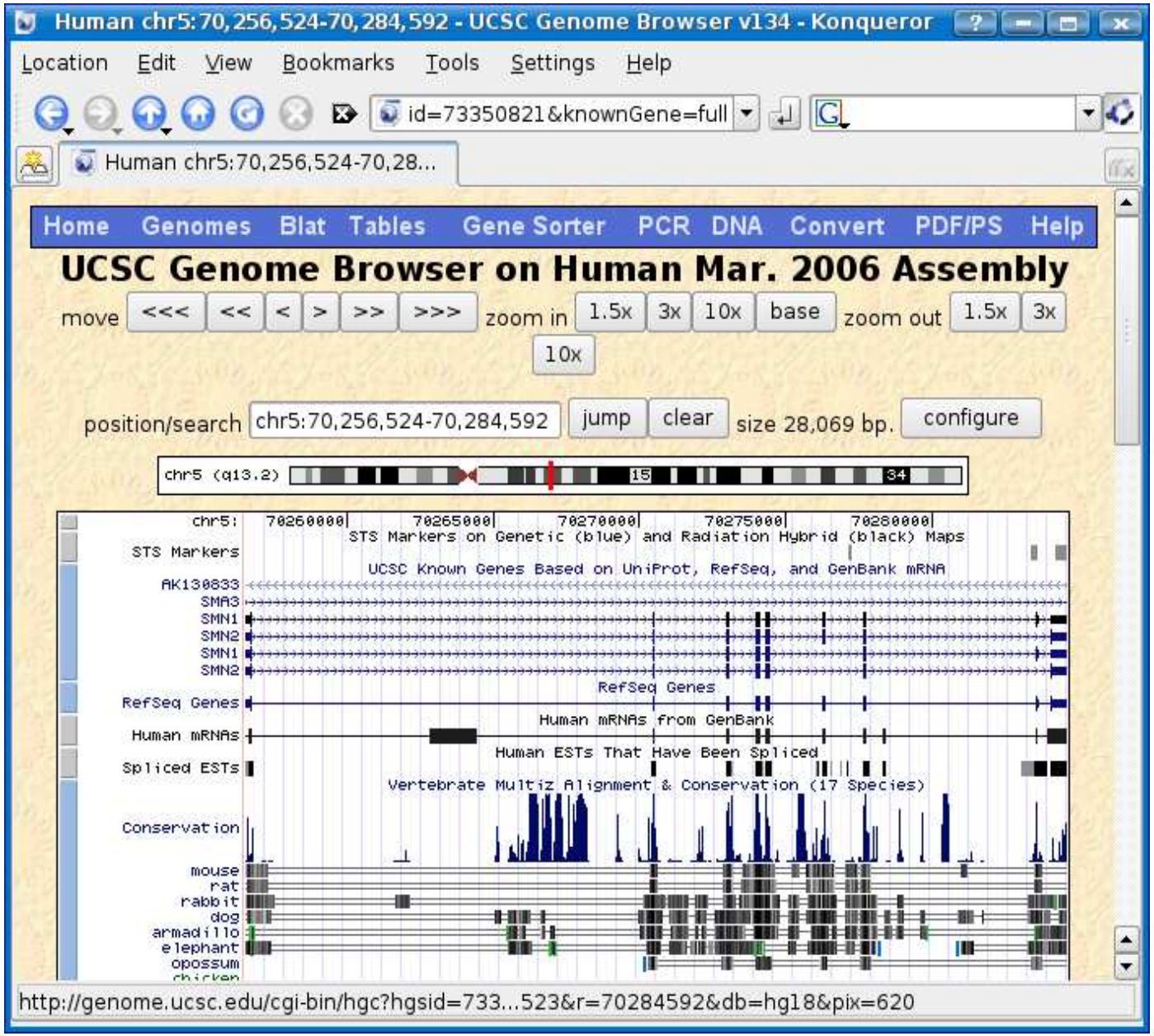
- Experimental Evidence Codes
  - IDA: Inferred from Direct Assay
  - IPI: Inferred from Physical Interaction
  - IMP: Inferred from Mutant Phenotype
  - IGI: Inferred from Genetic Interaction
  - IEP: Inferred from Expression Pattern
- Computational Analysis Evidence Codes
  - ISS: Inferred from Sequence or Structural Similarity
  - IGC: Inferred from Genomic Context
  - RCA: inferred from Reviewed Computational Analysis
- Author Statement Evidence Codes
  - TAS: Traceable Author Statement
  - NAS: Non-traceable Author Statement
- Curator Statement Evidence Codes
  - IC: Inferred by Curator
  - ND: No biological Data available
- Automatically-assigned Evidence Codes
  - IEA: Inferred from Electronic Annotation
- Obsolete Evidence Codes

Metabolické dráhy

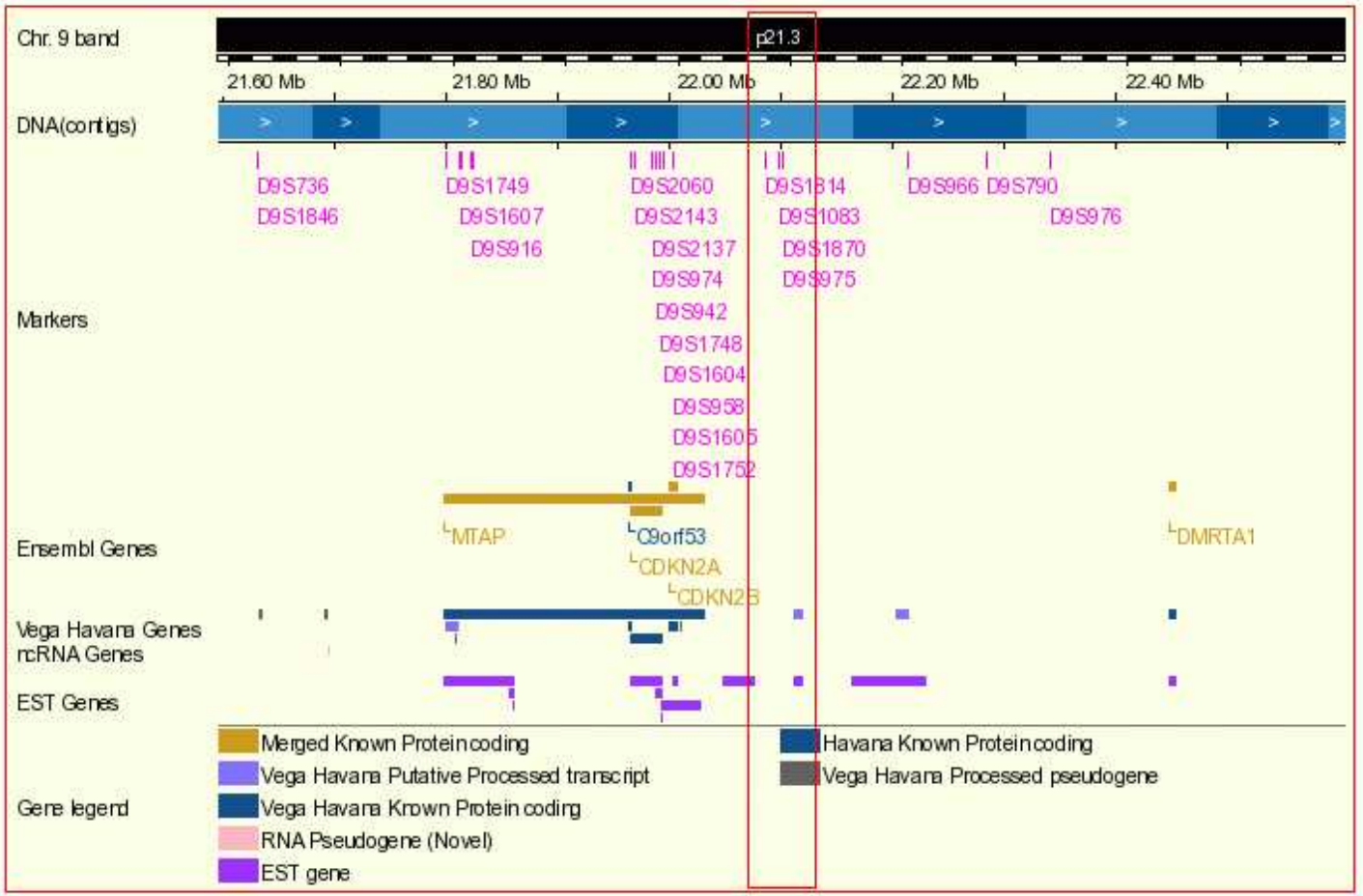


http://www.genome.jp/kegg/

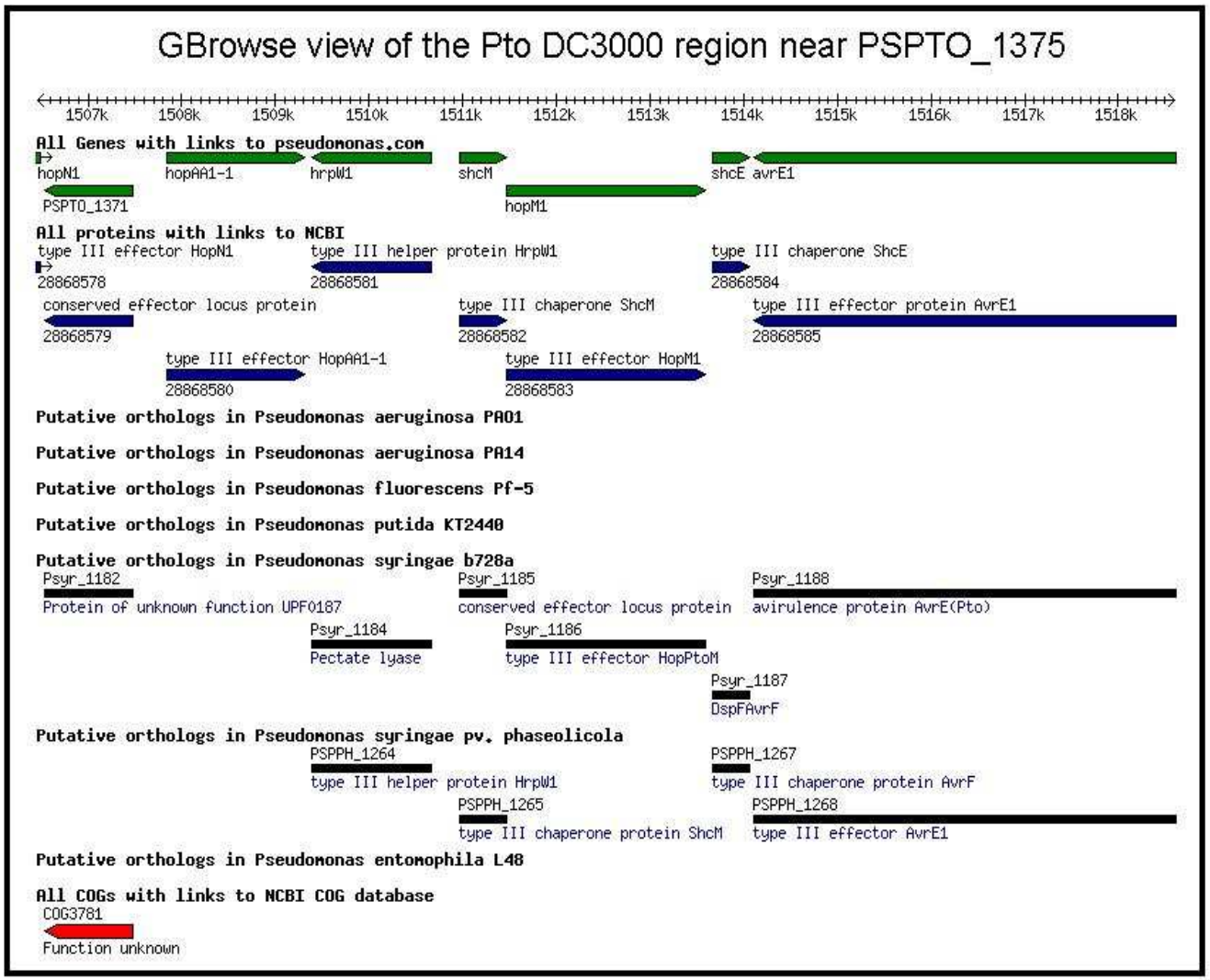
UCSC Genome Browser



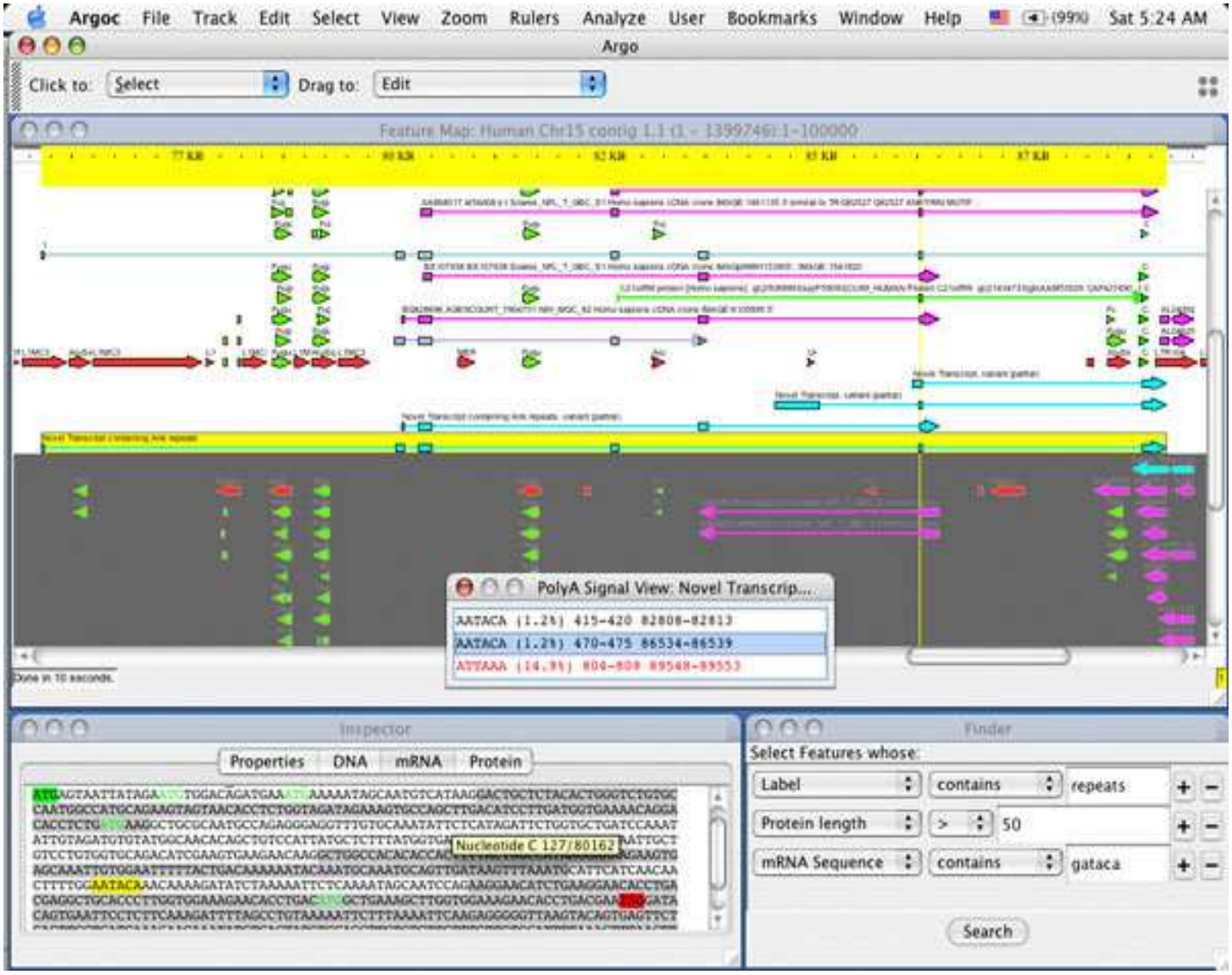
Ensembl Genome Browser



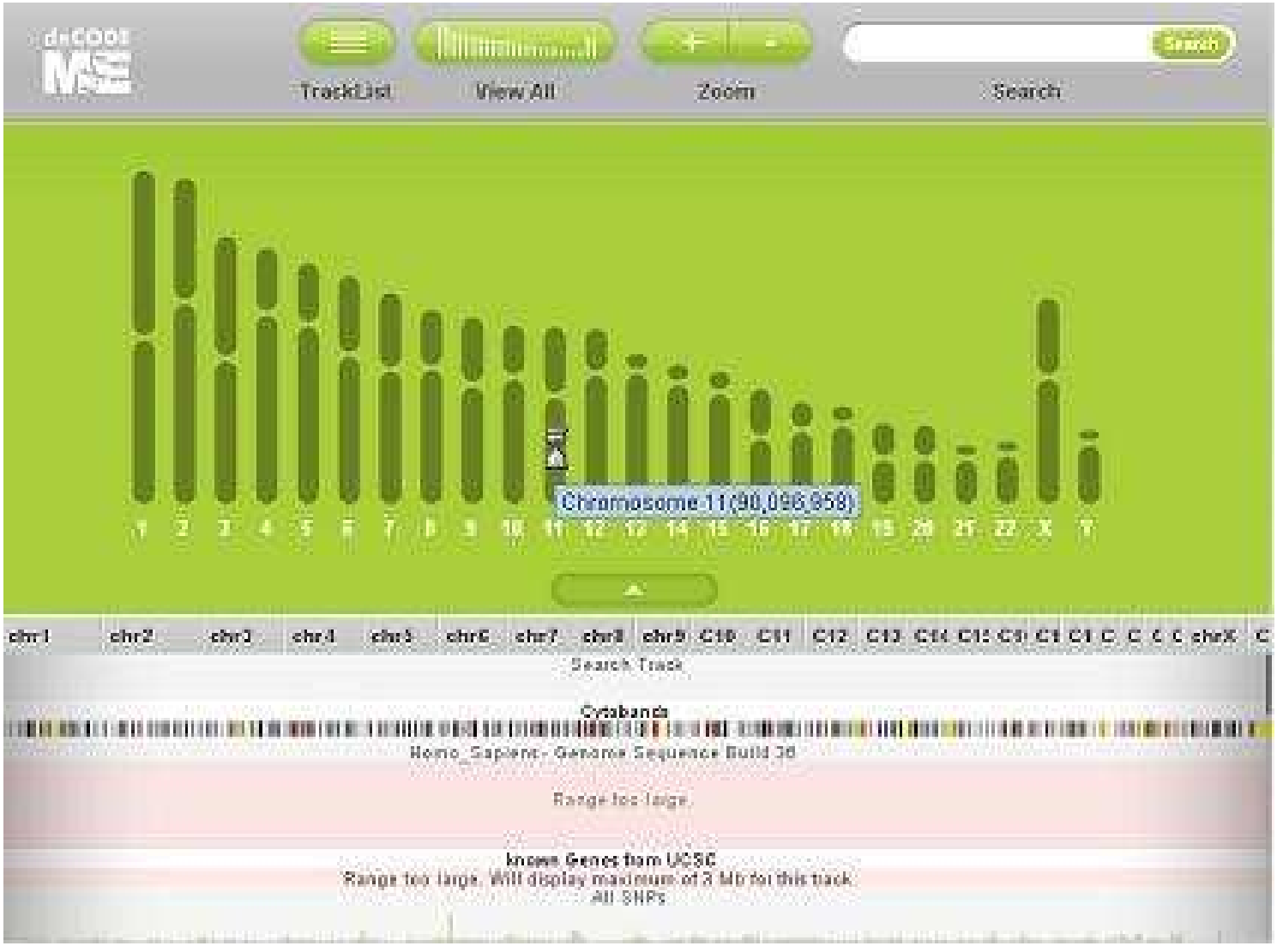
GBrowse


GBrowse view of the Pto DC3000 region near PSPTO_1375
1507k 1508k 1509k 1510k 1511k 1512k 1513k 1514k 1515k 1516k 1517k 1518k
All Genes with links to pseudomonas.com
hopN1
hopAA1-1
hrpW1
shcM
shcE avrE1
PSPTO_1371
hopM1
All proteins with links to NCBI
type III effector HopN1
type III helper protein HrpW1
type III chaperone ShcE
28868578
28868581
28868584
conserved effector locus protein
type III chaperone ShcM
type III effector protein AvrE1
28868579
28868582
28868585
type III effector HopAA1-1
type III effector HopM1
28868580
28868583
Putative orthologs in Pseudomonas aeruginosa PAO1
Putative orthologs in Pseudomonas aeruginosa PA14
Putative orthologs in Pseudomonas fluorescens Pf-5
Putative orthologs in Pseudomonas putida KT2440
Putative orthologs in Pseudomonas syringae b728a
Psyr_1182
Psyr_1185
Psyr_1188
Protein of unknown function UPF0187
conserved effector locus protein
avirulence protein AvrE(Pto)
Psyr_1184
Psyr_1186
Pectate lyase
type III effector HopPtoM
Psyr_1187
DspFAvrF
Putative orthologs in Pseudomonas syringae pv. phaseolicola
PSPPH_1264
PSPPH_1267
type III helper protein HrpW1
type III chaperone protein AvrF
PSPPH_1265
PSPPH_1268
type III chaperone protein ShcM
type III effector AvrE1
Putative orthologs in Pseudomonas entomophila L48
All COGs with links to NCBI COG database
COG3781
Function unknown

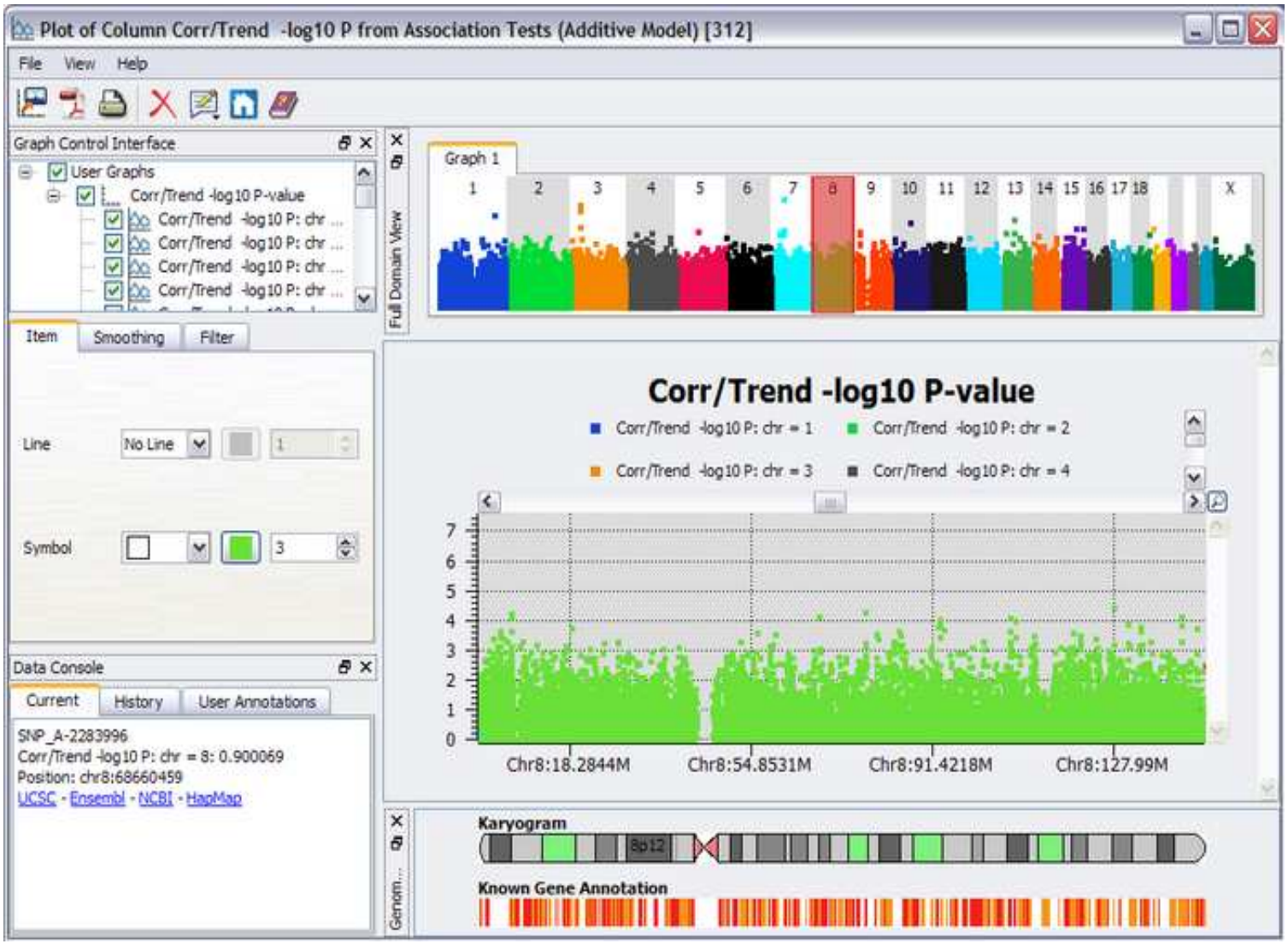
Argo


Argoc File Track Edit Select View Zoom Rulers Analyze User Bookmarks Window Help Sat 5:24 AM
Argo
Click to: Select
Drag to: Edit
Feature Map: Human Chr15 contig 1.1 (1 - 1399746):1-100000
PolyA Signal View: Novel Transcrip...
AATACA (1.2%) 415-420 82808-82813
AATACA (1.2%) 470-475 86534-86539
ATTAAA (14.3%) 804-809 89548-89553
Inspector
Properties DNA mRNA Protein
Nucleotide C 127/80162
Finder
Select Features whose:
Label contains repeats
Protein length > 50
mRNA Sequence contains gataca
Search

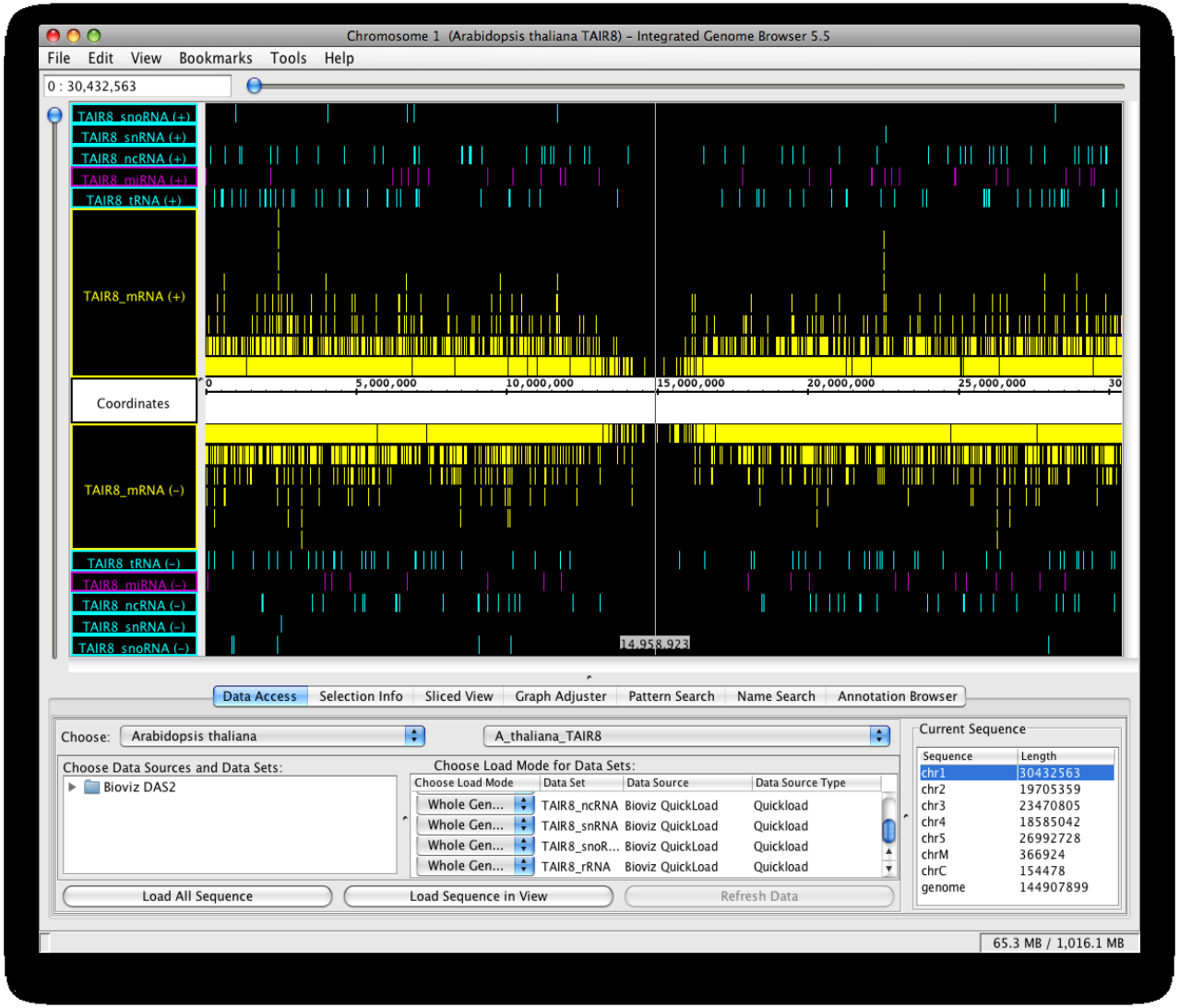
DecodeMe Browser



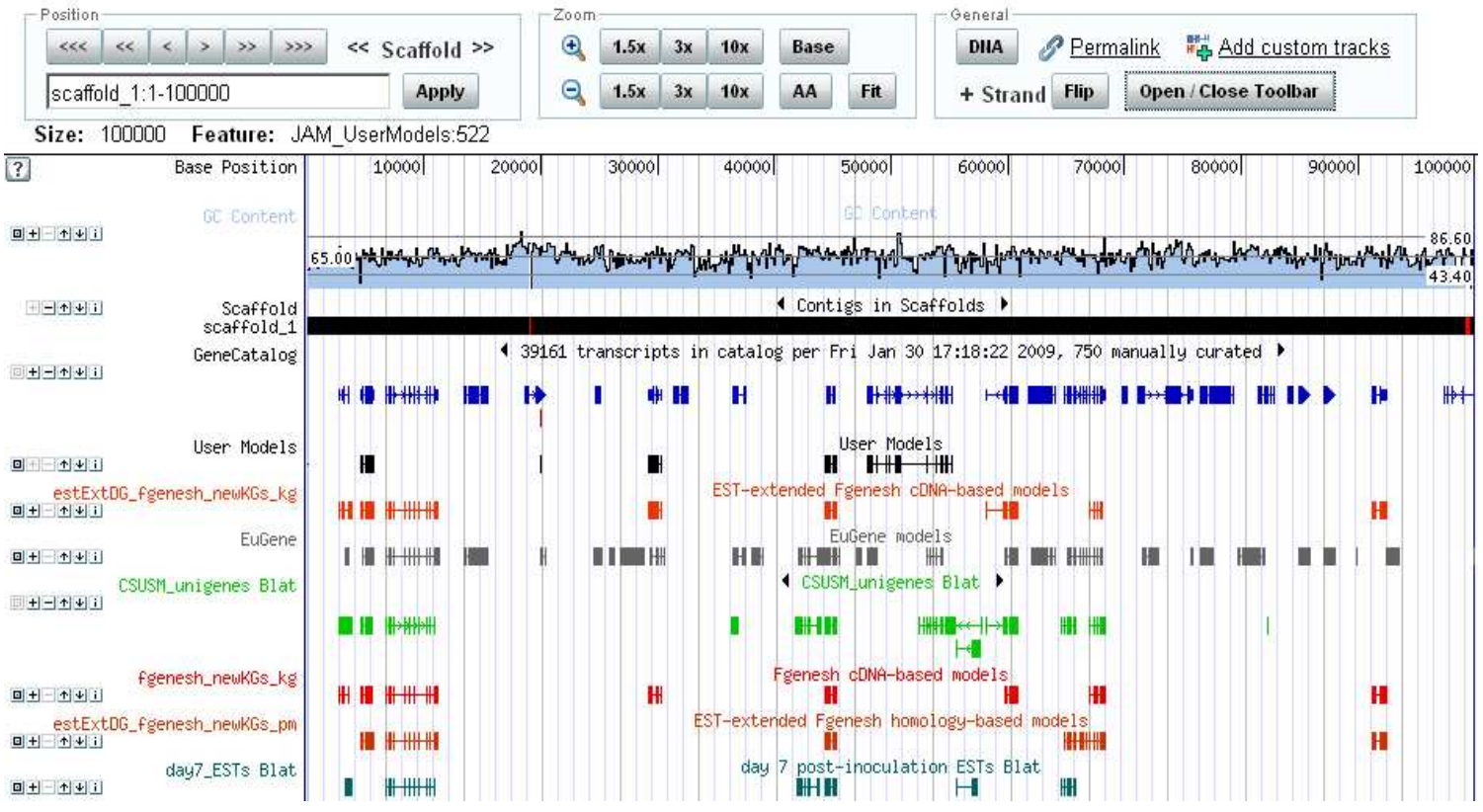
Golden Helix Genome Browser



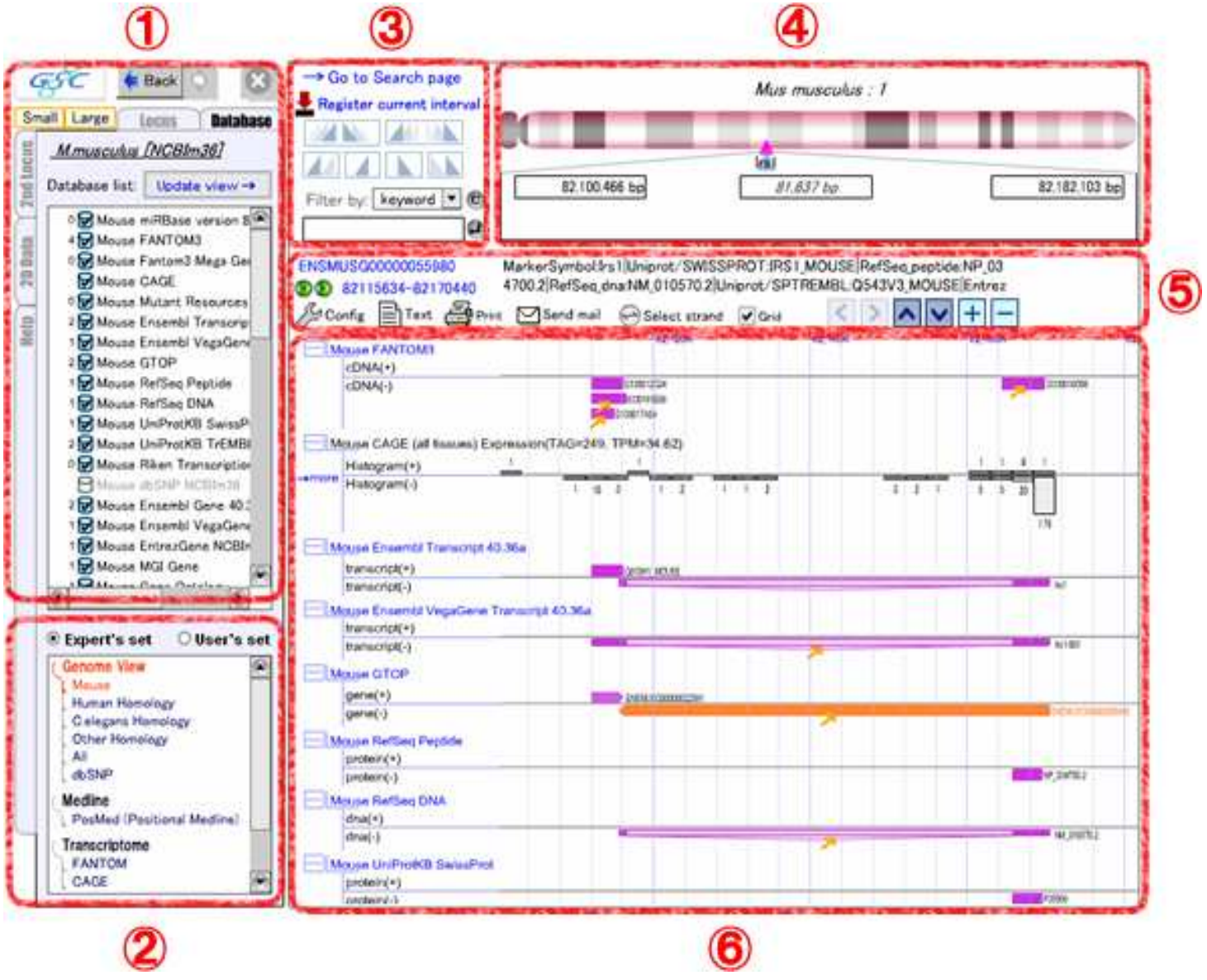
IGB


Chromosome 1 (Arabidopsis thaliana TAIR8) - Integrated Genome Browser 5.5
File Edit View Bookmarks Tools Help
0 : 30,432,563
TAIR8_snoRNA (+)
TAIR8_snRNA (+)
TAIR8_ncRNA (+)
TAIR8_miRNA (+)
TAIR8_tRNA (+)
TAIR8_mRNA (+)
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30
Coordinates
TAIR8_mRNA (-)
TAIR8_tRNA (-)
TAIR8_miRNA (-)
TAIR8_ncRNA (-)
TAIR8_snRNA (-)
TAIR8_snoRNA (-)
14,958,923
Data Access
Selection Info
Sliced View
Graph Adjuster
Pattern Search
Name Search
Annotation Browser
Choose:
Arabidopsis thaliana
A_thaliana_TAIR8
Choose Data Sources and Data Sets:
Bioviz DAS2
Choose Load Mode for Data Sets:
Choose Load Mode
Data Set
Data Source
Data Source Type
Whole Gen...
TAIR8_ncRNA
Bioviz QuickLoad
Quickload
Whole Gen...
TAIR8_snRNA
Bioviz QuickLoad
Quickload
Whole Gen...
TAIR8_snoR...
Bioviz QuickLoad
Quickload
Whole Gen...
TAIR8_rRNA
Bioviz QuickLoad
Quickload
Load All Sequence
Load Sequence in View
Refresh Data
Current Sequence
Sequence
Length
chr1 30432563
chr2 19705359
chr3 23470805
chr4 18585042
chr5 26992728
chrM 366924
chrC 154478
genome 144907899
65.3 MB / 1,016.1 MB

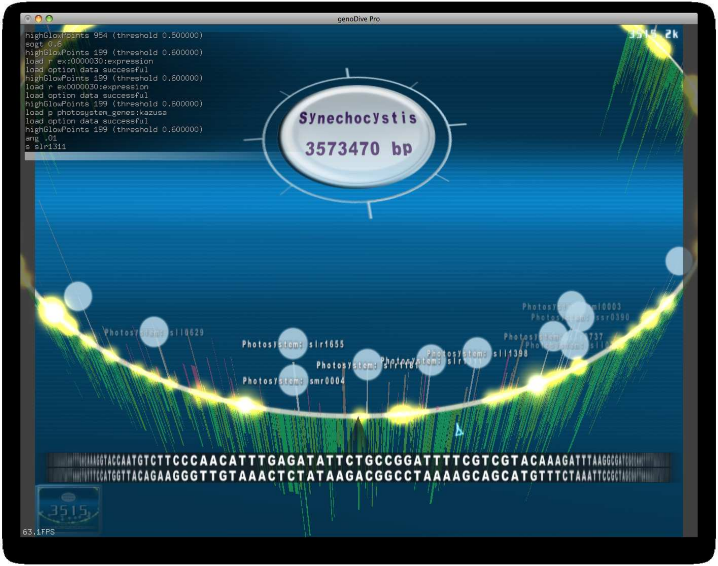
JGI Browser


Position
<<< << < > >> >>>
<< Scaffold >>
scaffold_1:1-100000
Apply
Zoom
1.5x 3x 10x Base
1.5x 3x 10x AA Fit
General
DNA
Permalink
Add custom tracks
+ Strand
Flip
Open / Close Toolbar
Size: 100000 Feature: JAM_UserModels:522
Base Position
10000 20000 30000 40000 50000 60000 70000 80000 90000 100000
GC Content
GC Content
65.00
86.60
43.40
Scaffold
scaffold_1
Contigs in Scaffolds
GeneCatalog
39161 transcripts in catalog per Fri Jan 30 17:18:22 2009, 750 manually curated
User Models
User Models
estExtDG_fgenesh_newKGs_kg
EST-extended Fgenesh cDNA-based models
EuGene
EuGene models
CSUSM_unigenes Blat
CSUSM_unigenes Blat
fgenesh_newKGs_kg
Fgenesh cDNA-based models
estExtDG_fgenesh_newKGs_pm
EST-extended Fgenesh homology-based models
day7_ESTs Blat
day 7 post-inoculation ESTs Blat

RIKEN Genome Browser



GenoDive


genoDive Pro
highGlowPoints 954 (threshold 0.500000)
sogt 0.6
highGlowPoints 199 (threshold 0.600000)
load r ex:0000030:expression
load option data successful
highGlowPoints 199 (threshold 0.600000)
load r ex0000030:expression
load option data successful
highGlowPoints 199 (threshold 0.600000)
load p photosystem_genes:kazusa
load option data successful
highGlowPoints 199 (threshold 0.600000)
ang .01
s slr1311
Synechocystis
3573470 bp
Photosystem: slr1655
Photosystem: smr0004
63.1FPS

Příště

Analýza proteinových sekvencí, strukturních a funkčních dat

Outline

For Further Reading

X